NONPARAMETRIC STUDY

Uber Pickups in NYC

Group 19

Xiaoou Su

Abstract

Rising standards of living have led to significant changes in traffic volume during holiday periods, especially in New York City. Uber, as a transportation network company, founded in 2009 and launched UberPool in New York City in 2014, served about 110 million worldwide users in 2019. However, recently, Uber has stopped accepting new drivers on their respective platforms in New York City. The challenge of traffic has been illustrated during holiday periods. Are Uber passengers mostly tourists? In this study, we use data from Kaggle which contains data on over 4.5 million Uber pickups in New York City from April to September 2014, and 14.3 million more Uber pickups from January to June 2015. We discuss different the pickups pattern over different time period and the nonparametric tests result showed that Uber passengers take rides mostly for daily purposes.

Keywords: Nonparametric; Uber; Traffic

Original Data

The data (figure 1.) consists four variables, Date/Time, Latitude, Longitude and Base. There are five bases included, each base associated with the company issued the licenses. (figure 2.)

Date/Time	Lat	Lon	Base
4/1/14 0:11	40.769	-73.9549	B02512
4/1/14 0:17	40.7267	-74.0345	B02512
4/1/14 0:21	40.7316	-73.9873	B02512
4/1/14 0:28	40.7588	-73.9776	B02512
4/1/14 0:33	40.7594	-73.9722	B02512
4/1/14 0:33	40.7383	-74.0403	B02512
4/1/14 0:39	40.7223	-73.9887	B02512
4/1/14 0:45	40.762	-73.979	B02512
4/1/14 0:55	40.7524	-73.996	B02512

Figure 1.

Base Code	Base Name
B02512	Unter
B02598	Hinter
B02617	Weiter
B02682	Schmecken
B02764	Danach-NY

Figure 2.

Objectives and Methods

1. **Monthly Differences**
   Whether the monthly pickup in New York City has identical distribution from April to June in 2014
   Method: **Kruskal-Wallis Test**
2. **Rush Hour Differences**
   Whether there is a significant relationship between weekdays (1=weekdays, 0=weekends) and hours (1=8am-8pm, 0=other)
   Method: **Chi-Square Test**
3. **Rush Hour Differences Over Months**
   Whether weekdays (1=weekdays, 0=weekends) and hours (1=8am-8pm, 0=other) are independent within each month (1=April, 2=May, 3=June)
   Method: **Mantel-Haenszel Test**
4. **Yearly Differences**
   Whether the monthly pickup in New York City has identical distribution in 2014 and 2015
   Methods: **Kolmogorov-Smirnov Test / Wilcoxon Rank Sum Test**
5. **Weekday Differences in School Area**
   Whether the daily pickup in School Area has identical distribution (Tuesday and Saturday)
   Methods: **Kolmogorov-Smirnov Test / Wilcoxon Rank Sum Test**

Data

In order to conduct nonparametric tests, we processed the original data and added 5 extra variables based on Date/Time. (figure 1.)

Date.Time <fctr>	Lat <dbl>	Lon <dbl>	Base <fctr>	month <chr>	weekday <chr>	hour <int>	is.weekday <dbl>	is.rush <dbl>
4/1/2014 0:11:00	40.7690	-73.9549	B02512	April	Tuesday	0	1	0
4/1/2014 0:17:00	40.7267	-74.0345	B02512	April	Tuesday	0	1	0
4/1/2014 0:21:00	40.7316	-73.9873	B02512	April	Tuesday	0	1	0
4/1/2014 0:28:00	40.7588	-73.9776	B02512	April	Tuesday	0	1	0
4/1/2014 0:33:00	40.7594	-73.9722	B02512	April	Tuesday	0	1	0
4/1/2014 0:33:00	40.7383	-74.0403	B02512	April	Tuesday	0	1	0
4/1/2014 0:39:00	40.7223	-73.9887	B02512	April	Tuesday	0	1	0
4/1/2014 0:45:00	40.7620	-73.9790	B02512	April	Tuesday	0	1	0
4/1/2014 0:55:00	40.7524	-73.9960	B02512	April	Tuesday	0	1	0
4/1/2014 1:01:00	40.7575	-73.9846	B02512	April	Tuesday	1	1	0

Figure 3. Processed Data (Month, Weekday, Hour: 0-23, Is.weekday: Weekday = 1, Weekends = 0, Is.rush: Rush hour (8am-8pm) = 1, Other = 0)

Analysis and Results

1. **Kruskal-Wallis Test** (Monthly differences for pickups in 2014)
   Formula: Test statistic: H

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_{i\cdot} - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}$$

   Description:
   We collected total pickups (figure 4.) in April, May, June for five bases in 2014 and conducted Kruskal-Wallis Test to test whether there are significant differences from April to June.

	B02512	B02598	B02617	B02682	B02764
April	35536	183263	108001	227808	9908
May	36765	260549	122734	222883	9504
June	32509	242975	184460	194926	8974

Figure 4.

   Results:
   We cannot reject the null hypothesis that they have identical distributions. (figure 5.)

```
        Kruskal-Wallis rank sum test

data:  kw_pickups by kw_month
Kruskal-Wallis chi-squared = 0.18, df = 2, p-value = 0.9139
```

Figure 5.

   Conclusion:
   It is different from our initial guess that for the summer time increasing number of tourists will increase some rides. The conclusion is passengers are "regular" riders (not tourists)

2. **Chi-Square Test** (Rush hour difference in weekdays and weekends in April 2014)
   Formula: Test statistic: X^2

$$X^2 = \sum_{i=1}^{k} \frac{(x_i - m_i)^2}{m_i}$$

   Description:
   We constructed contingency table (figure 6.) for weekdays and rush hour in April 2014 (Weekdays = 1, Weekends = 0; Rush (8am-8pm) = 1, Others = 0)

	Rush(=1)	Other(=0)
Weekdays(=1)	306532	129515
Weekends(=0)	81079	47390

Figure 6.

   Results:
   We reject the null hypothesis. (figure 7.) Weekdays and rush hour are dependent.

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  tb1
X-squared = 2381.3, df = 1, p-value < 2.2e-16
```

Figure 7.

Conclusion:
There are more pickups in rush hour during weekdays. We concluded that passengers may take Uber for daily rides

3. **Mantel-Haenszel Test** (Weekdays, hours | months)
Formula: Test statistic: MH

$$\mathrm{MH}-\chi^2_{\mathrm{C}}=\frac{\left(\left|\sum_j A_j-\sum_j E_{A_j}\right|-0.5\right)^2}{\sum_j V_{A_j}},$$

Description:
We constructed contingency tables (figure 8.) for weekdays and rush hour in April, May and June 2014.

```
, ,  = April

   0      1
0  47390  81079
1 129515 306532

, ,  = May

      0      1
0  54971 104187
1 147179 346098

, ,  = June

      0      1
0  57588 103432
1 152326 350498
```

Figure 8.

Results:
Chi-squared = 5155.46; P-value ≈ 0. We reject the null hypothesis. Weekdays and rush hour are dependent conditionally in months.
Conclusion:
Even for the summer time the relationship between weekdays and rush hour remains. We conclude most Uber riders are “regular” passengers. (not tourists).

4. **Kolmogorov-Smirnov Test / Wilcoxon Rank Sum Test** (yearly differences in 2014 and 2015, April)
Formula: Test statistic: Kolmogorov-Smirnov: D

$$D_{n,m}=\sup_x|F_{1,n}(x)-F_{2,m}(x)|$$

Test statistic: Wilcoxon Rank Sum: U

$$U_1=R_1-\frac{n_1(n_1+1)}{2}$$

Description:
We collected total pickups (figure 9.) in April for five bases in 2014 and 2015 and conducted Kolmogorov-Smirnov Test / Wilcoxon Rank Sum Test to test whether there are significant pickup differences between 2014 and 2015

	B02512	B02598	B02617	B02682	B02764
2014	35536	183263	108001	227808	9908
2015	14388	243488	326930	544313	919556

Figure 9.

Results:
Pickups for 2014 and 2015 are not identically distributed with 0.10 confidence level. (figure 10.)

```
        Wilcoxon rank sum test

data:  x1 and x2
W = 21, p-value = 0.09524
alternative hypothesis: true location shift is not equal to 0
```

```
        Two-sample Kolmogorov-Smirnov test

data:  x1 and x2
D = 0.8, p-value = 0.07937
alternative hypothesis: two-sided
```

Figure 10.

Conclusion:
There are more pickups in 2015 than 2014

5. **Kolmogorov-Smirnov Test / Wilcoxon Rank Sum Test** (weekday difference in April 2014 in school area)
Formula: Test statistic: Kolmogorov-Smirnov: D

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

Test statistic: Wilcoxon Rank Sum: U

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$$

Description:
We collected total pickups (figure 11.) in school area (figure 11.) for five bases in 2014 and conducted Kolmogorov-Smirnov Test / Wilcoxon Rank Sum Test to test whether there are significant pickup differences between Tuesday and Saturday

	B02512	B02598	B02617	B02682	B02764
Tuesday	3	27	21	31	1
Saturday	8	27	11	27	1

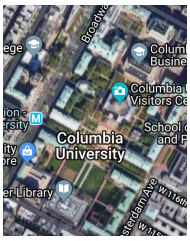

Figure 11. (Pickups; School Area)

Results:
We cannot reject null hypothesis. Pickups for Tuesday and Saturday are identically distributed. (figure 12.)

```
        Two-sample Kolmogorov-Smirnov test

data:  x2 and x6
D^- = 0.2, p-value = 0.8187
alternative hypothesis: the CDF of x lies below that of y
```

```
        Wilcoxon rank sum test with continuity correction

data:  x2 and x6
W = 13.5, p-value = 0.9155
alternative hypothesis: true location shift is not equal to 0
```

Figure 12.

Conclusion:
There are not more pickups during weekends than weekdays in school area. We concluded that students may not take rides for daily purpose.

Conclusion

It is different from our initial guess that increasing number of tourists increases number of Uber pickups in New York City during summer time. The nonparametric tests showed that passengers of Uber may take rides as daily purpose. Specifically, there is no monthly pickup difference; there are more pickups in rush hour during weekdays; even for the summer time the relationship between weekdays and rush hour remains. We concluded that passengers may take Uber for daily rides. The last two tests show there are more pickups in 2015 than 2014 and there are not more pickups during weekends than weekdays in school area. We concluded that Uber Pool launched in 2014 increased pickups and students may not take rides for daily purpose.